

REPORT DOCUMENTATION PAGE			Form Approved OMB NO. 0704-0188		
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY)		2. REPORT TYPE Technical Report		3. DATES COVERED (From - To) -	
4. TITLE AND SUBTITLE Confidence sets for network structure			5a. CONTRACT NUMBER W911NF-11-1-0036		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER 611103		
6. AUTHORS Edoardo Airoldi, David Choi, Patrick Wolfe			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES Harvard University Office of Sponsored Research 1350 Massachusetts Ave. Holyoke 727 Cambridge, MA 02138 -			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211			10. SPONSOR/MONITOR'S ACRONYM(S) ARO		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) 58153-MA-MUR.1		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.					
14. ABSTRACT Latent variable models are frequently used to identify structure in dichotomous network data, in part because they give rise to a Bernoulli product likelihood that is both well understood and consistent with the notion of exchangeable random graphs. In this article we propose conservative confidence sets that hold with respect to these underlying Bernoulli parameters as a function of any given partition of network nodes, enabling us to assess estimates of 'residual' network structure, that is, structure that cannot be explained by known covariates and thus					
15. SUBJECT TERMS network structure, latent variable, stochastic blockmodel					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Patrick Wolfe
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU			19b. TELEPHONE NUMBER 617-496-1448

Report Title

Confidence sets for network structure

ABSTRACT

Latent variable models are frequently used to identify structure in dichotomous network data, in part because they give rise to a Bernoulli product likelihood that is both well understood and consistent with the notion of exchangeable random graphs. In this article we propose conservative confidence sets that hold with respect to these underlying Bernoulli parameters as a function of any given partition of network nodes, enabling us to assess estimates of 'residual' network structure, that is, structure that cannot be explained by known covariates and thus cannot be easily verified by manual inspection. We demonstrate the proposed methodology by analyzing student friendship networks from the National Longitudinal Survey of Adolescent Health that include race, gender, and school year as covariates. We employ a stochastic expectation-maximization algorithm to fit a logistic regression model that includes these explanatory variables as well as a latent stochastic blockmodel component and additional node-specific effects. Although maximum-likelihood estimates do not appear consistent in this context, we are able to evaluate confidence sets as a function of different blockmodel partitions, which enables us to qualitatively assess the significance of estimated residual network structure relative to a baseline, which models covariates but lacks block structure.

Confidence sets for network structure

Edoardo M. Airolidi, David S. Choi, and Patrick J. Wolfe

Harvard University, Cambridge, MA 02138, USA

(airolidi@stat.harvard.edu, dchoi@seas.harvard.edu, patrick@seas.harvard.edu)

Abstract

Latent variable models are frequently used to identify structure in dichotomous network data, in part because they give rise to a Bernoulli product likelihood that is both well understood and consistent with the notion of exchangeable random graphs. In this article we propose conservative confidence sets that hold with respect to these underlying Bernoulli parameters as a function of any given partition of network nodes, enabling us to assess estimates of *residual* network structure, that is, structure that cannot be explained by known covariates and thus cannot be easily verified by manual inspection. We demonstrate the proposed methodology by analyzing student friendship networks from the National Longitudinal Survey of Adolescent Health that include race, gender, and school year as covariates. We employ a stochastic expectation-maximization algorithm to fit a logistic regression model that includes these explanatory variables as well as a latent stochastic blockmodel component and additional node-specific effects. Although maximum-likelihood estimates do not appear consistent in this context, we are able to evaluate confidence sets as a function of different blockmodel partitions, which enables us to qualitatively assess the significance of estimated residual network structure relative to a baseline, which models covariates but lacks block structure.

1 Introduction

Network datasets comprising *edge* measurements $A_{ij} \in \{0, 1\}$ of a binary, symmetric, and anti-reflexive relation on a set of n nodes, $1 \leq i < j \leq n$, are fast becoming of paramount interest in the statistical analysis and data mining literatures (Goldenberg et al., 2010). A common aim of many models for such data is to test for and explain the presence of network *structure*, primary examples being *communities* and *blocks* of nodes that are equivalent in some formal sense. Algorithmic formulations of this problem take varied forms and span many literatures, touching on subjects as diverse as spectral graph theory (Chung, 1996), statistical physics (Albert and Barabasi, 2002; Newman, 2003), theoretical computer science (Cooper and Frieze, 2003; Achlioptas et al., 2009;

Easley and Kleinberg, 2010), economics (Jackson, 2008), and social network analysis (Wasserman and Faust, 1994).

One popular modeling assumption for network data is to assume dyadic independence of the edge measurements when conditioned on a set of latent variables (Snijders and Nowicki, 1997; Handcock et al., 2007; Airoldi et al., 2008; Hoff, 2008). The number of latent parameters in such models generally increases with the size of the graph, however, meaning that computationally intensive fitting algorithms may be required and standard consistency results may not always hold. As a result, it can often be difficult to assess statistical significance or quantify the uncertainty associated with parameter estimates. This issue is evident in literatures focused on community detection, where common practice is to examine whether algorithmically identified communities agree with prior knowledge or intuition (Newman, 2006; Traud et al., 2011). This practice is less useful if additional confirmatory information is unavailable, or if detailed uncertainty quantification is desired.

Confidence sets are a standard statistical tool for uncertainty quantification, but they are not yet well developed for network data (Wasserman and Faust, 1994). In this paper, we propose a family of confidence sets for network structure that apply under the assumption of a Bernoulli product likelihood. The form of these sets stems from a stochastic blockmodel formulation which reflects the notion of latent nodal classes, and they provide a new tool for the analysis of estimated or algorithmically determined network structure. We demonstrate usage of the confidence sets by analyzing a sample of 26 adolescent friendship networks from the National Longitudinal Survey of Adolescent Health (available at <http://www.cpc.unc.edu/addhealth>), using a *baseline* model that only includes explanatory covariates and heterogeneity in the nodal degrees. We employ these confidence sets to validate departures from this baseline model taking the form of *residual* community structure. Though the confidence sets we employ are known to be conservative (Choi et al., 2011), we show that they are effective in identifying putative residual structure in these friendship network data.

2 Model specification and inference

We represent dichotomous network data via a sociomatrix $A \in \{0, 1\}^{N \times N}$ that reflects the adjacency structure of a simple, undirected graph on N nodes. In keeping with the latent variable network analysis literature, we assume entries $\{A_{ij}\}$ for $i < j$ to be independent Bernoulli random variables with associated success probabilities $\{P_{ij}\}_{i < j}$, and complete A as a symmetric matrix with zeros

along its main diagonal. The corresponding data log-likelihood is given by

$$L(A; P) = \sum_{i < j} A_{ij} \log(P_{ij}) + (1 - A_{ij}) \log(1 - P_{ij}), \quad (1)$$

where each P_{ij} can itself be modeled as a function of latent as well as explanatory variables. Typically, the nodal indices themselves impart no inferential information and the latent variable formulation is exchangeable; in fact, in lieu of independence one may start with a much weaker assumption that the graph itself is generated from an exchangeable distribution (see [Bickel and Chen \(2009\)](#); [Hoff \(2009\)](#)). Exchangeability accommodates many types of dependencies that are of interest in network analysis, such as triadic closure.

Given an instantiation of A and a latent variable model for the probabilities $\{P_{ij}\}_{i < j}$, it is natural to seek a quantification of the uncertainty associated with estimates of these Bernoulli parameters. A standard approach in non-graphical settings is to posit a parametric form for each P_{ij} and then compute confidence intervals with respect to the corresponding parameter values, for example by appealing to standard maximum-likelihood asymptotics. However, as mentioned earlier, the formulation of most latent variable network models dictates an increasing number of parameters as the number of network nodes grows; as we have observed in our own exploratory data analysis, this amount of expressive power appears necessary to capture many salient characteristics of network data. As a result, standard asymptotic results do not necessarily apply, leaving open questions for inference.

2.1 A logistic regression model for network structure

To illustrate the complexities that can quickly arise in this inferential setting, we adopt a latent variable network model with a standard flavor: a logistic regression model that simultaneously incorporates aspects of blockmodels, additive effects, and explanatory variables (see [Hoff \(2009\)](#) for a more general formulation). Specifically, we incorporate a K -class stochastic blockmodel component parameterized in terms of a symmetric matrix $\Theta \in \mathbb{R}^{K \times K}$ and a membership vector $z \in \{1, \dots, K\}^N$ whose values denote the class of each node, with P_{ij} depending on $\Theta_{z_i z_j}$. A vector of additional node-specific latent variables α is included to account for heterogeneity in the observed nodal degrees, along with a vector of regression coefficients β corresponding to explanatory variables $x(i, j)$. Thus we obtain the log-odds parameterization

$$\log \frac{P_{ij}}{1 - P_{ij}} = \Theta_{z_i z_j} + \alpha_i + \alpha_j + x(i, j)' \beta, \quad (2)$$

where we further enforce the identifiability constraint that $\sum_i \alpha_i = 0$.

Intuitively, the blockmodel component of model (2) serves to enforce stochastic equivalence of nodes in the same class (Holland et al., 1983; Fienberg et al., 1985), after accounting for node-specific additive effects and covariates. Blockmodeling in this manner is historically motivated by the deterministic notion of structural equivalence (Lorrain and White, 1971; Doreian et al., 2005), under which the network must remain identical under any relabeling of nodes within each community. While this notion is typically too strong to hold exactly, relaxations such as regular equivalence¹ and other variations (Doreian et al., 2005) can be thought of as deterministic precursors to the stochastic blockmodel (Nowicki and Snijders, 2001) and its modern extensions (Airoldi et al., 2005, 2008; Xing et al., 2010).

2.2 Likelihood-based inference

Analogous to the deterministic setting, exact maximization of the corresponding log-likelihood $L(A; z, \Theta, \alpha, \beta, x)$ is computationally demanding even for moderately large K and N , owing to the total number of possible nodal partitions induced by z . However, when z is fixed the corresponding conditional likelihood maximization task is convex, and so we may adopt a stochastic expectation-maximization (EM) approach to inference in which z is treated as missing data (Dempster et al., 1977; van Dyk and Meng, 2010).

Algorithm 1 Stochastic Expectation-Maximization Fitting of model (2)

1. Set $t = 0$ and initialize $(z^{(0)}, \Theta^{(0)}, \alpha^{(0)}, \beta^{(0)})$.
 2. For iteration t , do:

E-step Sample $z^{(t)} \propto \exp\{L(z | A; \Theta^{(t)}, \alpha^{(t)}, \beta^{(t)}, x)\}$
(e.g., via Gibbs sampling)

M-step Set $(\Theta^{(t)}, \alpha^{(t)}, \beta^{(t)}) = \operatorname{argmax}_{\Theta, \alpha, \beta} L(\Theta, \alpha, \beta | z^{(t)}; A, x)$
(convex optimization)
 3. Set $t \leftarrow t + 1$ and return to Step 2.
-

Exact sampling of z according to the E-step of Algorithm 1 requires computing the normalizing constant of $\exp\{L(z | A; \Theta, \alpha, \beta, x)\}$, and so we may instead approximate this step by a series of univariate draws from a coordinate-wise Gibbs sampler that conditions on all other elements of z in turn. When K is moderate, both steps of Algorithm 1 are computationally tractable, since the Gibbs sampler requires only discrete distributions over K values, and the M-step optimization problem is convex. We note that the confidence sets developed in Section 2.3 will accommodate

¹In deterministic equivalence, the bipartite subgraph of edges between two communities is either empty or complete. In regular equivalence, the same subgraph is either empty, or has no isolated nodes. See (Doreian et al., 2005) for further discussion.

the sub-optimality of Algorithm 1.

When α and β are fixed to zero, model (2) reduces to a re-parameterization of the standard stochastic blockmodel. Consistency results for this model have been developed for a range of conditions (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001; Bickel and Chen, 2009; Rohe et al., 2011; Choi et al., 2011; Celisse et al., 2011); alternative fitting methods with lower computational complexity, such as spectral clustering (Rohe et al., 2011) and modularity maximization (Bickel and Chen, 2009), are also available for the stochastic blockmodel, with (Bickel and Chen, 2009) noting that such approaches appear to perform well in practice. However, it is not clear how uncertainty in z and Θ should be quantified or even concisely expressed: in this vein, previous efforts to assess the robustness of fitted structure include (Karrer et al., 2008), in which community partitions are analyzed under perturbations of the network, and (Massen and Doye, 2006), in which the behavior of local minima resulting from simulated annealing is examined; a likelihood-based test is proposed in (Copic et al., 2009) to compare sets of community divisions.

Without the blockmodel components z and Θ , the model of Eq. (2) reduces to a generalized linear model whose likelihood can be maximized by standard methods. If α is further constrained to equal 0, the model is finite dimensional and standard asymptotic results for inference can be applied. Otherwise, the increasing dimensionality of α brings consistency into question, and in fact certain negative results are known for the related p_1 exponential random graph model (Holland and Leinhardt, 1981). Specifically, (Haberman, 1981) reports that the maximum likelihood estimator for the p_1 model exhibits bias with magnitude equal to its variance. Although estimation error does converge asymptotically to zero for the p_1 model, it is generally not known how to generate confidence intervals or hypothesis tests; (Wasserman and Weaver, 1985) prescribes reporting standard errors only as summary statistics, with no association to p -values. The predictions of (Haberman, 1981) were replicated (reported below) when fitting simulated data drawn from the model of Eq. (2) with parameters matched to observed characteristics of the Adolescent Health friendship networks.

2.3 Confidence sets for network structure

Instead of quantifying the uncertainty associated with estimates of the model parameters $(z, \Theta, \alpha, \beta)$, we directly find confidence sets for the Bernoulli likelihood parameters $\{P_{ij}\}_{i < j}$. To this end, for any fixed K and class assignment z , define symmetric matrices $\bar{\Phi}, \hat{\Phi}$ in $[0, 1]^{K \times K}$ element-wise for $1 \leq a \leq b$ as

$$\bar{\Phi}_{ab}^{(z)} = \frac{1}{n_{ab}} \sum_{i < j} P_{ij} 1\{z_i = a, z_j = b\}, \quad \hat{\Phi}_{ab}^{(z)} = \frac{1}{n_{ab}} \sum_{i < j} A_{ij} 1\{z_i = a, z_j = b\},$$

with n_{ab} denoting the maximum number of possible edges between classes a and b (i.e., the corresponding number of Bernoulli trials). Thus $\bar{\Phi}_{ab}^{(z)}$ is the expected proportion of edges between (or within, if $a = b$) classes a and b , under class assignment z , and $\hat{\Phi}_{ab}^{(z)}$ is its corresponding sample proportion estimator.

Intuitively, $\bar{\Phi}^{(z)}$ measures *assortativity* by z ; whenever the sociomatrix A is unstructured, elements of $\bar{\Phi}^{(z)}$ should be nearly uniform for any choice of partition z . When strong community structure is present in A , however, these elements should instead be well separated for corresponding values of z . Thus, it is of interest to examine a confidence set that relates $\hat{\Phi}_{ab}^{(z)}$ to its expected value $\bar{\Phi}_{ab}^{(z)}$ for a range of partitions z . To this end, we may define such a set by considering a weighted sum of the form $\sum_{a \leq b} n_{ab} D(\hat{\Phi}_{ab}^{(z)} || \bar{\Phi}_{ab}^{(z)})$, where $D(p || p') = p \log(p/p') + (1-p) \log[(1-p)/(1-p')]$ denotes the (nonnegative) Kullback–Leibler divergence of a Bernoulli random variable with parameter p' from that of one with parameter p .

A confidence set is then obtainable via direct application of the following theorem.

Theorem 1 (Choi et al., 2011). *Let $\{A_{ij}\}_{i < j}$ be comprised of $\binom{N}{2}$ independent Bernoulli(P_{ij}) trials, and let $\mathcal{Z} = \{1, \dots, K\}^N$. Then with probability at least $1 - \delta$,*

$$\sup_{z \in \mathcal{Z}} \sum_{a \leq b} n_{ab} D(\hat{\Phi}_{ab}^{(z)} || \bar{\Phi}_{ab}^{(z)}) \leq N \log K + (K^2 + K) \log \left(\frac{N}{K} + 1 \right) + \log \frac{1}{\delta}. \quad (3)$$

Because Eq. (3) holds uniformly over all class assignments, we may choose to apply it directly to the value of z obtained from Algorithm 1—and because it does not assume any particular form of latent structure, we are able to avoid the difficulties associated with determining confidence sets directly for the parameters of latent variable models such as Eq. (2). However, it is important to note that this generality comes at a price: In simulation studies undertaken in (Choi et al., 2011) as well as those detailed below, we have observed the bound of Eq. (3) to be loose by a multiplicative factor ranging from 3 to 7 on average.

2.4 Estimator consistency and confidence sets

Recalling our above discussion of estimator consistency for the related p_1 model, we undertook a small simulation study to investigate the consistency of maximum-likelihood (ML) estimation in a “baseline” version of model (2) with $K = 1$ and the corresponding (scalar) value of Θ set equal to zero. We compared estimates for the cases $\alpha = 0$ versus α unconstrained for 500 graphs generated randomly from a model of the form specified in Eq. (2) based on school 8 of the Add-Health data set. The number of nodes $N = 204$ and covariates $x(i, j)$ matched that of School 8 in the Adolescent Health friendship network dataset, and the regression coefficient vector $\beta = (-2.6, 0.025, 0.9, -1.6)'$,

Element of β	$\Theta = 0, \alpha = 0$	$\Theta = 0$ only
Intercept	-0.001 (0.004)	2.26 (0.070)
Gender	0.003 (0.004)	-0.005 (0.004)
Race	-0.001 (0.004)	-0.03 (0.005)
Grade	0.006 (0.003)	0.04 (0.003)

Table 1: Empirical bias (with standard errors) of ML-estimated components of β under a baseline model, for the cases $\alpha = 0$ versus α unconstrained. Note the change in estimated bias when α is included in the model.

set to match the ML estimate of β for School 8, fitted via logistic regression with $\Theta = 0, \alpha = 0$. The covariates $x(i, j)$ comprised of an intercept term, an indicator for whether students i and j shared the same gender, an indicator for shared race, and their difference in school grade.

The inclusion of α in the model of Eq. (2) appears to give rise to a loss of estimator consistency, as shown in Table 1 where the empirical bias of each component of β is reported. This suggests, as we alluded to above, that inferential conclusions based on parameter estimates from latent variable models should be interpreted with caution.

To explore the tightness of the confidence sets given by the bound in Eq. (3), we fitted the full model specified in Eq. (2) with K in the range 2–6 to 50 draws from a restricted version of the model corresponding to each of the 26 schools in our dataset. In the same manner described above, each simulated graph shared the same size and covariates as its corresponding school in the dataset, with β fixed to its ML-fitted value with $\Theta = 0, \alpha = 0$. The empirical divergence term $\sum_{a \leq b} n_{ab} D(\hat{\Phi}_{ab}^{(z)} || \bar{\Phi}_{ab}^{(z)})$ under the approximate ML partition determined via Algorithm 1 was then tabulated for each of these 1300 fits, and compared to its 95% confidence upper bound given by Eq. (3). The empirical divergences are reported in the histogram of Fig. 1 as a fraction of the upper bound.

It may be seen from Fig. 1 that the largest divergence observed was less than 41% of its corresponding bound, with 95% of all divergences less than 22% of their corresponding bound. This analysis verifies that, despite the issues associated with estimator consistency in the context of latent variable models, our confidence sets can be used as a tool to assess putative network structure. This analysis also provides an indication of how inflated the confidence set sizes are expected to be in practice; while conservative in nature, they seem usable for practical situations.

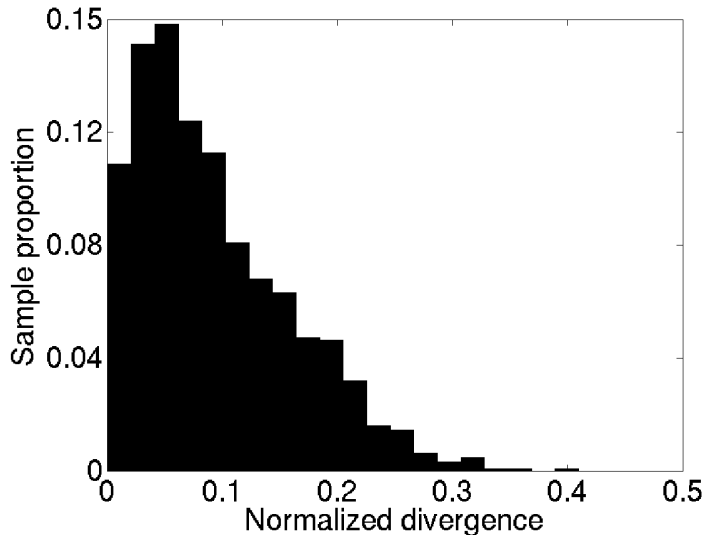


Figure 1: Divergence terms $\sum_{a \leq b} n_{ab} D(\hat{\Phi}_{ab}^{(z)} || \bar{\Phi}_{ab}^{(z)})$ as fractions of 95% confidence set values, shown for approximate maximum-likelihood fits to 1300 randomly graphs matched to the 26-school friendship network dataset. Smaller values indicate slack in the upper bound of Eq. (3).

3 Analysis of adolescent health networks

We now turn our attention to an analysis of friendship networks from the National Longitudinal Survey of Adolescent Health; see, e.g., (Goodreau et al., 2009) for a recent analysis and in-depth discussion. Previous studies have looked for trends which held across several schools, such as clustering by race (González et al., 2007). Within a school, however, it may be desirable to go beyond reported covariates such as gender, race, or grade in describing social structure. Here we examine the schools individually to find residual block structure not explained by the covariates. Since we will be unable to verify such blocks by checking against explanatory variables, we rely on the confidence sets developed above to assess significance of the discovered block structure.

Recall from Section 2.3 that Eq. (3) enables us to calculate confidence sets with respect to Bernoulli parameters $\{P_{ij}\}$ for any class membership vector z in terms of the corresponding sample proportion matrices $\hat{\Phi}^{(z)}$. Then, by comparing values of $\hat{\Phi}^{(z)}$ to a baseline model obtained by fitting $K = 1, \Theta = 0$ (thus removing the stochastic block component from Eq. (2)), we may evaluate whether or not the observed sample counts are consistent with the structure predicted by the baseline model. This procedure provides a kind of notional p -value to qualitatively assess significance of the residual structure induced by any choice of z .

In Section 3.1 below, we describe the dataset. Then, in Section 3.2, we verify empirically that

School	Nodes	Edges	School	Nodes	Edges	School	Nodes	Edges
1	69	220	21	377	1531	67	456	926
2	105	349	22	614	2450	70	76	344
3	32	91	26	551	2066	71	74	358
5	157	730	29	569	2534	72	352	1398
6	108	378	38	521	1925	76	43	185
8	204	809	55	336	803	77	25	106
9	248	1004	56	446	1394	78	432	1334
10	678	2795	63	98	283	80	594	1745
18	284	1189	66	644	2865			

Table 2: Network characteristics (number of students and reported friendship edges) corresponding to each of the 26 schools used for analysis.

the models provide reasonable fits and conclusions, first by fitting our baseline model ($K = 1$) and checking its behavior, and then by fitting a “pure” stochastic blockmodel ($K > 1, \alpha = 0, \beta = 0$) and observing that we can recover block structure corresponding to held-out grade covariates. In Section 3.3, we consider estimates of z obtained via Algorithm 1 for K in the range 2–6, and use our notion of confidence sets introduced above to evaluate how unlikely the corresponding observed configurations of $\hat{\Phi}^{(z)}$ are, relative to an underlying set of Bernoulli random variables whose parameters are assumed to be equal to those fitted under our baseline model.

3.1 Network data description

The data we examine were collected in 1994–95 through an in-school survey administered to 90,118 students at 145 schools. In the 26-school subset of the data we consider, each student was given a paper-and-pencil questionnaire and a copy of a roster listing every student in that school; the schools that we considered did not have a *sister school* from where friends could also be listed (which would complicate the analysis) and they had fewer than 700 students. Students were asked to list up to five friends of each gender, and whether or not they had interacted with these friends at particular venues within a certain time period. Gender, race (5 categories), and school year (grades 7–12) covariates were also recorded for each student. The number of students and reported friendships for this subset of schools are listed in Table 2.

We treat the friendship network data from these 26 schools as undirected graphs, with an edge present between two students whenever either of them indicated any type of friendship with the other. Other approaches are also possible and can be expected to influence inferential results (Butts, 2003), though we verified through exploratory data analysis that estimates were qualitatively similar even if sparser versions of these networks were constructed by instead requiring mutual indication of friendship within student pairs, resulting in a roughly 3-fold decrease in the observed number of

edges in our data (cf. Table 2). As in Section 2.4, for each student pair the corresponding vector $x(i, j)$ of explanatory variables comprised an intercept term, binary indicator variables indicating shared gender and shared race, and an ordinal value equal to the absolute difference in the grades of the students; model (2) was then fitted to these data, with results described below.

3.2 Model checking

We first fit model (2) with $\Theta = 0$ and $\alpha = 0$, since it reduces to a logistic regression with explanatory variables $x(i, j)$, for which standard asymptotic results apply. The parameter fits were examined and an analysis of deviance was conducted. The fits were observed to be well behaved in this setting; estimates of β and their corresponding standard errors indicate a clustering effect by grade that is stronger than that of either shared gender or race. The fitting thus appears consistent with the previous analysis of (Handcock et al., 2007), in which student friendship clusters were observed to closely align with grades. An analysis of deviance, where each variable was withheld from the model, resulted in similar conclusions: Average deviances across the 26 schools were -69 , -238 , and -3760 for gender, race, and grade respectively, with p -values below 0.05 indicating significance in all but 3, 7, and 0 of the schools for each of the respective covariates; these schools had small numbers of students, with a maximum N of 108.

Because standard analysis techniques do not apply to the parameters α and Θ , before fitting the full model we first fit models using these terms separately, and checked if the results were intuitive. When α was re-introduced into the model of Eq. (2), its components were observed to correlate highly with the sequence of observed nodal degrees in the data, as expected. (Recall that consistency results are not known for this model, so that p -values cannot be associated with deviances or standard errors; however, in our simulations the maximum-likelihood estimates showed moderate errors, as discussed in Section 2.4.) For two of the schools, the resulting model was degenerate, whereas for the remaining schools the α -degree correlation had a range of 0.78–0.94 and a median value of 0.89. Estimates of β did not undergo qualitative significant changes from their earlier values when the restriction $\alpha = 0$ was lifted.

A “pure” stochastic blockmodel ($\alpha = 0, \beta = 0$) was fitted to our data over the range $K \in \{2, \dots, 6\}$, to observe if the resulting block structure replicates that of any of the known covariates. Figure 2 shows counts of students by latent class (under the approximate maximum-likelihood estimate of z) and grade for School 6; it can be seen that the recovered grouping of students by latent class is closely aligned with the grade covariate, particularly for grades 7–10.

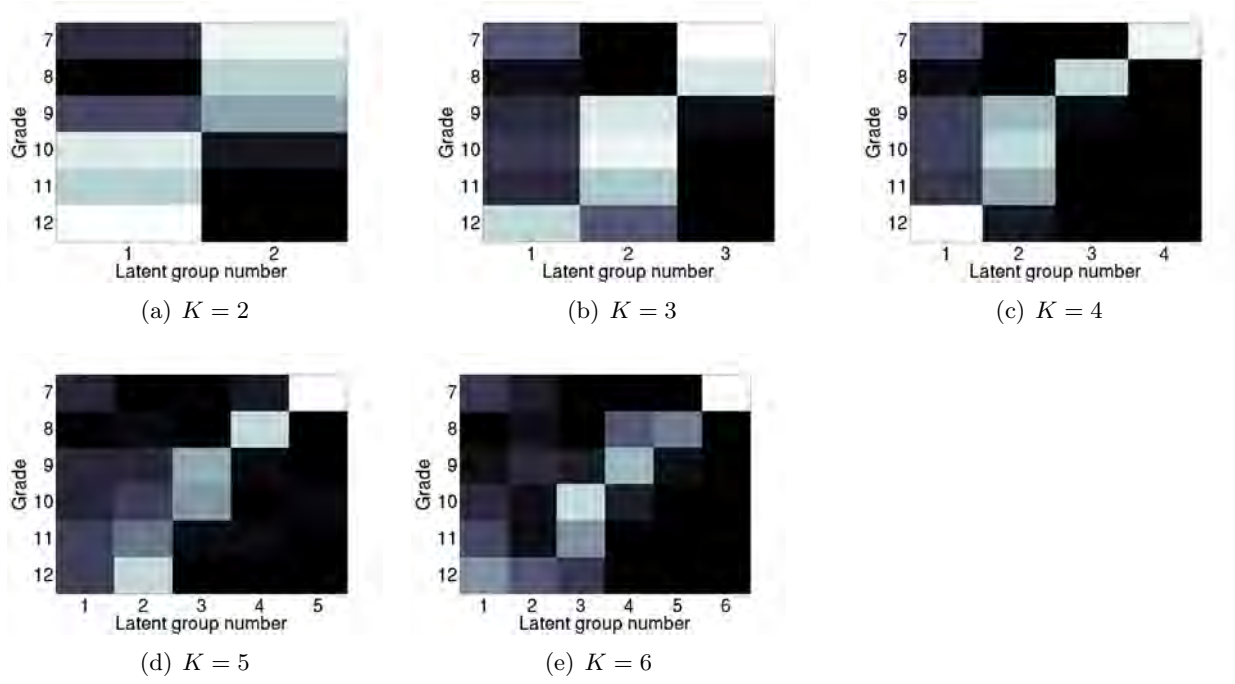


Figure 2: Student counts resulting from a stochastic blockmodel fit for $K \in \{2, \dots, 6\}$, arranged by latent block and school year (grade) for School 6. The inferred block structure approximately aligns with the grade covariate (which was not included in this model).

3.3 Residual block structure

We now report on the assessment of residual block structure in the Adolescent Health friendship network data. Recalling that the confidence sets obtained with Eq. (3) hold uniformly for all partitions of equal size, independently of how they are chosen, we therefore may freely modify the fitting procedure of Algorithm 1 to obtain partitions that exhibit the greatest degree of structure. Bearing in mind the high observed α -degree correlation as discussed above, we replaced the latent variable vector α in the model of Eq. (2) with a more parsimonious categorical covariate determined by grouping the observed network degrees according to the ranges 0–3, 4–7, and 8– ∞ . We also expanded the covariates by giving each race and grade pairing its own indicator function. These modifications would be inappropriate for the baseline model, as dyadic independence conditioned on the covariates would be lost, and standard errors for β would be larger; however, the changes were useful for improving the output of Algorithm 1 without invalidating Eq. (3).

Fig. 3 depicts partitions for which the observed $\hat{\Phi}^{(z)}$, fitted for various $K > 1$ using the modified version of Algorithm 1 detailed above, is “far” from its nominal value under the baseline model fitted with $K = 1$, in the sense that the corresponding 95% Bonferroni-corrected confidence set bound is exceeded. We observe that in each partition, the number of apparently visible communities exceeds

School	K	Div. (Bound)	Jaccard coefficient / Variance ratio			
			Gender	Race	Grade	Degree
10	6	0.0064 (0.0062)	0.14	0.16	0.097	0.93
18	5	0.0150 (0.0150)	0.17	0.19	0.14	0.88
21	6	0.0140 (0.0120)	0.15	0.16	0.12	0.95
22	5	0.0064 (0.0061)	0.18	0.14	0.11	0.99
26	3	0.0049 (0.0045)	0.25	0.21	0.13	0.99
29	6	0.0091 (0.0075)	0.15	0.16	0.10	0.88
38	5	0.0073 (0.0073)	0.17	0.18	0.17	0.86
55	4	0.0100 (0.0100)	0.20	0.18	0.21	0.97
56	6	0.0120 (0.0099)	0.15	0.14	0.15	0.98
66	6	0.0069 (0.0066)	0.15	0.16	0.099	0.91
67	3	0.0055 (0.0055)	0.25	0.23	0.25	1.00
72	4	0.0099 (0.0095)	0.21	0.21	0.12	0.96
78	6	0.0100 (0.0100)	0.15	0.12	0.15	0.98
80	4	0.0054 (0.0053)	0.20	0.19	0.15	0.99

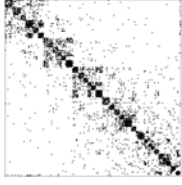
Table 3: Block structure assessments corresponding to Fig. 3. Small Jaccard coefficient values and variance ratios approaching 1 indicate a lack of alignment with covariates and hence the identification of residual structure in the corresponding partition.

K , and they are comprised of small numbers of students. This effect is due to the intersection of grade and z -induced clustering.

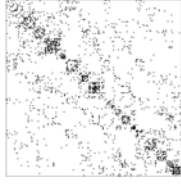
We take as our definition of nominal value the quantity $\hat{\Phi}^{(z)}$ computed under the baseline model, which we denote by $\Phi^{(z)}$. Table 3 lists normalized divergence terms $\binom{N}{2}^{-1} \sum_{a \leq b} n_{ab} D(\hat{\Phi}_{ab}^{(z)} || \Phi_{ab}^{(z)})$, Bonferroni-corrected 95% confidence bounds, and measures of alignment between the corresponding partitions z and the explanatory variables. The alignment with the covariates are small, as measured by the Jaccard similarity coefficient and ratio of within-class to total variance², signifying the residual quality of the partitions, while the relatively large divergence terms signify that the Bonferroni-corrected confidence set bounds for each school have been met or exceeded.

We note that the usage of covariate information was necessary to detect small student groups; without the incorporation of grade effects, we would require a much larger value of K for Algorithm 1 to detect the observed network structure (a concern noted by (Hoff, 2008) in the absence of covariates), which in turn would inflate the confidence set, leading to an inability to validate the observed structure from that predicted by a baseline model.

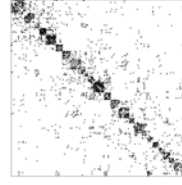
²The alignment scores are defined as follows. The Jaccard similarity coefficient is defined as $|A \cap B| / |A \cup B|$, where $A, B \subset \binom{N}{2}$ are the student pairings sharing the same latent class or the same covariate value, respectively. See (Traud et al., 2011) for further network-related discussion. Variance ratio denotes the within-class degree variance divided by the total variance, averaged over all classes.



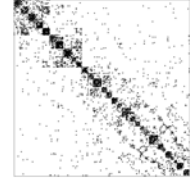
(a) School 10, $K = 6$



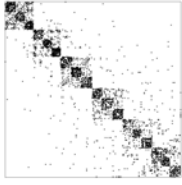
(b) School 18, $K = 5$



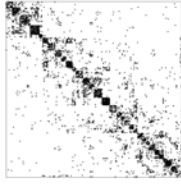
(c) School 21, $K = 6$



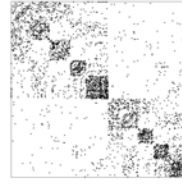
(d) School 22, $K = 5$



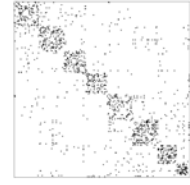
(e) School 26, $K = 3$



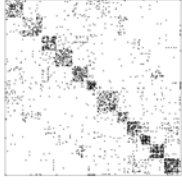
(f) School 29, $K = 6$



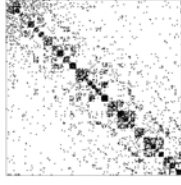
(g) School 38, $K = 5$



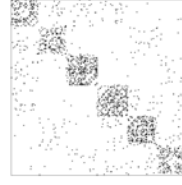
(h) School 55, $K = 4$



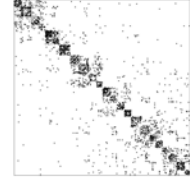
(i) School 56, $K = 6$



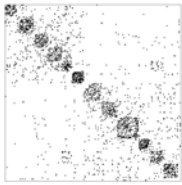
(j) School 66, $K = 6$



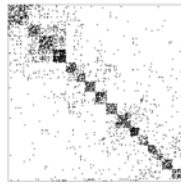
(k) School 67, $K = 3$



(l) School 72, $K = 4$



(m) School 78, $K = 6$



(n) School 80, $K = 4$

Figure 3: Adjacency matrices for schools exhibiting residual block structure as described in Section 3.3, with nodes ordered by grade and corresponding latent classes.

4 Concluding remarks

In this article we have developed confidence sets for assessing inferred network structure, by leveraging our result derived in Choi et al. (2011). We explored the use of these confidence sets with an application to the analysis of Adolescent Health survey data comprising friendship networks from 26 schools.

Our methodology can be summarized as follows. In lieu of a parametric model, we assume dyadic independence with Bernoulli parameters $\{P_{ji}\}$. We introduced a baseline model ($K = 1$) that incorporates degree and covariate effects, without block structure. Algorithm 1 was then used to find highly assortative partitions of students which are also far from partitions induced by the explanatory covariates in the baseline model. Differences in assortativity were quantified by an empirical divergence statistic, which was compared to an upper bound computed from Eq. (3) to check for significance and to generate confidence sets for $\{P_{ji}\}$. While the upper bound in Eq. (3) is known to be loose, simulation results in Figure 1 suggest that the slack is moderate, leading to useful confidence sets in practice.

In our procedure, we cannot quantify the uncertainty associated with the estimated baseline model, since the nodal parameter estimates lack consistency. As a result, we cannot conduct a formal hypothesis test for $\Theta = 0$. However, for a baseline model where the MLE is known to be consistent, we conjecture that such a hypothesis test should be possible by incorporating the confidence set associated with the MLE.

Despite concerns regarding estimator consistency in this and other latent variable models, we were able to show that the notion of confidence sets may instead be used to provide a (conservative) measure of residual block structure. We note that many open questions remain, and are hopeful that this analysis may help to shed light on some important current issues facing practitioners and theorists alike in statistical network analysis.

Acknowledgments

Work supported in part by the National Science Foundation, National Institutes of Health, Army Research Office, and Office of Naval Research, USA. Additional funding was provided by the Harvard Medical School’s Milton Fund. This research uses data from Add Health, a project funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development. Information on how to obtain the Add Health data files is available online at <http://www.cpc.unc.edu/addhealth>. We thank the reviewers for their comments and suggestions that helped improve the paper.

References

- D. Achlioptas, R. M. D’Souza, and J. Spencer. Explosive percolation in random networks. *Science*, 323(5920):1453–1455, 2009.
- E. Airoldi, D. Blei, E. Xing, and S. Fienberg. A latent mixed membership model for relational data. In *Proceedings of the 3rd Int. Workshop on Link Discovery*, pages 82–89. ACM, 2005.
- E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic block-models. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.
- R. Albert and A. L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(47), 2002.
- P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl Acad. Sci. U.S.A.*, 106:21068–21073, Dec. 2009.
- C.T. Butts. Network inference, error, and informant (in) accuracy: a Bayesian approach. *Soc. Networks*, 25:103–140, May 2003.
- A. Celisse, J.J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. Arxiv preprint arXiv:1105.3288, 2011.
- D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with growing numbers of classes. *Biometrika*, 2011. to appear.
- F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, December 1996.
- C. Cooper and A. M. Frieze. A general model of web graphs. *Random Structures and Algorithms*, 22(3):311–335, 2003.
- J. Copic, M. O. Jackson, and A. Kirman. Identifying community structures from network data via maximum likelihood methods. *B.E. J. Theoretical Economics*, 9, Sept. 2009.
- A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
- P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized Blockmodeling*. Cambridge Univ. Pr., Cambridge, U.K., 2005.
- D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
- S. E. Fienberg, M. M. Meyer, and S. S. Wasserman. Statistical analysis of multiple sociometric relations. *J. Am. Stat. Assoc.*, 80:51–67, Mar. 1985.

- A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundation and Trends in Machine Learning*, 2(2):1–117, 2010.
- M.C. González, H.J. Herrmann, J. Kertész, and T. Vicsek. Community structure and ethnic preferences in school friendship networks. *Physica A.*, 379(1):307–316, 2007.
- S.M. Goodreau, J.A. Kitts, and M. Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46:103–125, Feb. 2009.
- SJ Haberman. Comment on Holland and Leinhardt. *J. Am. Stat. Assoc.*, 76:60–62, Mar. 1981.
- M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *J. R. Stat. Soc. A*, 170:301–354, Mar. 2007.
- P. D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Adv. in Neural Information Processing Systems*, pages 657–664. MIT Press, Cambridge, MA, 2008.
- P. D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational Math. Organization Theory*, 15:261–272, Dec. 2009.
- P. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: Some first steps. *Soc. Networks*, pages 109–137, June 1983.
- P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.*, 76:33–50, Mar. 1981.
- M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.
- B. Karrer, E. Levina, and MEJ Newman. Robustness of community structure in networks. *Phys. Rev. E*, 77:46119–46128, Apr. 2008.
- F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *J. Math. Sociol.*, 1:49–80, Mar. 1971.
- C.P. Massen and J.P.K. Doye. Thermodynamics of community structure. Arxiv preprint cond-mat/0610077, 2006.
- M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
- M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl Acad. Sci. U.S.A.*, 103:8577–8582, June 2006.

- K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, 96:1077–1087, Sept. 2001.
- K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block-model. *Ann. Stat.*, 2011. to appear.
- T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classif.*, 14:75–100, Jan. 1997.
- A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.*, 2011. to appear.
- D. A. van Dyk and X.-L. Meng. Cross-fertilizing strategies for better em mountain climbing and da field exploration: A graphical guide book. *Statistical Science*, 25(4):429–449, 2010.
- S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge Univ. Pr., Cambridge, U.K., 1994.
- S. Wasserman and S.O.L. Weaver. Statistical analysis of binary relational data: parameter estimation. *J. Math. Psychol.*, 29:406–427, Dec. 1985.
- E. P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat.*, 4:535–566, Dec. 2010.